Comparing Modes and Samples in Experiments
DRAFT *

Raymond Duch
Centre for Experimental Social Sciences
Nuffield College
University of Oxford
raymond.duch@nuffield.ox.ac.uk

Pablo Beramendi
Duke University
pablo.beramendi@duke.edu

Denise Laroze
Centre for Experimental Social Sciences
Universidad de Santiago de Chile
denise.laroze@cess.cl

May 30, 2017

*Paper prepared for presentation at the Annual Meeting of the Midwest Political Science Association, Chicago, April 6-9, 2017.

Abstract

We assess the internal and external validity of interactive experiments that are conducted with subjects both in the lab and online. The assessments are based on cheating experiments in which subjects earn real money, are subject to a tax (that is redistributed to other subjects), and can lie about their earnings. We identify the mode effects that affect internal validity by comparing treatment effects in lab and online versions of the experiments. Secondly, we compare the treatment effects across different online (MTurk and U.K. CESS Online) and lab subject pools to determine their effect on external validity. For the most part treatment effects are robust across mode and samples. Nevertheless, we find some features of both mode and subject pool that might affect experimental outcomes.

1 Introduction: Lab and Online Experiments

Experiments are a powerful tool for helping us understand how individuals make decisions. And many of the most important experiments conducted in the social sciences involve subjects making decisions that are affected by the choices taken by other subjects participating in the experiment. Social scientists have traditionally conducted these experiments in laboratory settings, typically with student subject pools, because they facilitated such interactive decision making.

There are good reasons to conduct experiments in a conventional lab setting, particularly when subjects interact with each other in real time (Morton and Williams, 2009). This is the case for public goods games in which subjects contribute to a common pool and are compensated depending on the size of contributions by other subjects in the game. Concerns about the internal validity should take priority over external validity (McDermott, 2002) and lab experiments are the gold standard for ensuring internal validity. Researchers have greater control of the environments in laboratory experiment than other forms of experiment, which helps to establish internal validity of any causal claims.

But there is a recognition that the results from any single lab experiment are fragile (Maniadis, Tufano and List, 2014). And while efforts to replicate lab experiments have been relatively successful (Camerer et al., 2016), we understand that there are features of the experimental data generation process that can enhance the robustness of experimental findings. This essay focuses on two aspects of experimental design that contribute to the robustness of estimated treatment effects.

There is little question that statistical power increases the robustness of estimated treatment effects. Camerer et al. (2016), for example, find a strong correlation between sample size and replicability of treatment effects from economics lab experiments. Our interest is evaluating design strategies that increase statistical power for experiments in which subjects take decisions that affect the choices of other participants (and vice-versa) in real time.

Secondly, there is considerable direct and indirect evidence that diversity of the subject

pool is related to the robustness of treatment effects. Treatment effects may interact with characteristics of laboratory subjects in which case a more diverse subject pool becomes important. Druckman and Kam (2011) argue that student convenience sample could bias the estimates of causal effects if the size of a treatment effect depends on a characteristic of convenience sample with little variance. Belot, Duch and Miller (2015) have compared the student convenience samples with non-student subjects and find that certainly with respect to games involving other-regarding preferences students can differ significantly from non-students.

Online experiments clearly offer such power and diversity but at a potential cost. As is the case with any experimental design we are concerned with whether the random assignment of online subjects to treatment and control meet the stringent assumption that "their potential outcome depends solely on whether the subject itself receives the treatment." Of particular concern here is the classic exclusion restriction: Are potential outcomes affected simply by the fact that treatments are delivered in an online environment? This essay compares subjects and treatment effects for different experimental modes in an attempt to answer this question. We focus on comparing lab and online treatment effects. Strictly speaking these should be identical. They are not. The essays indicates why they might not be identical and suggests how online experiments might be designed in order to minimize violation of the exclusion restriction.

There are two important methodological issues addressed in this essay. First, how do treatment effects vary across types of convenience subject pools – in particular, treatment effects associated with interactive decision making? Second, how do treatment effects in interactive decision making experiments differ across experimental modes – essentially online versus traditional lab?

There are attempts to identify the heterogeneity of treatment effects by running parallel experiments in different modes (e.g. Clifford and Jerit, 2014; Weigold, Weigold and Russell, 2013; Arechar, Gachter and Molleman, 2017) and with different subject pools (e.g. Berinsky,

Huber and Lenz, 2012; Mullinix et al., 2016; Coppock, 2016). These studies have explored the similarities and differences in the treatment effects from experiments conducted under different settings and environments using non-interactive experiments, typically survey experiments.

Our contribution is to focus on interactive experiments and we attempt to distinguish differences in treatment effects associated with the mode versus the subject pool. We disentangle the effects of modes of experimental administration and the effects of different subject pools by comparing experimental results from various modes and convenience samples.

To achieve these goals, we run interactive experiments in both traditional lab and online contexts. Our incentivised tax compliance experiments, build on Beramendi and Duch (2014) and Duch and Solaz (2016). The tax compliance experiments were initially designed and implemented in the Nuffield Centre for Experimental Social Sciences (CESS) lab.

Results from the lab are compared to similar online tax compliance experiments. The online experiments are all administered by the Nuffield CESS. There were six different versions of the online tax compliance experiments: a quasi-interactive version conducted with M-Turk subjects; a quasi-interactive version conducted with the CESS UK online subject pool; a quasi-interactive version conducted with subjects from the CESS lab pool; an interactive version conducted with M-Turk subjects; an interactive version conducted with CESS UK online subjects; and an interactive online version conducted with subjects from the CESS lab pool.

2 Subject Pool versus Experimental Modes Effects

Subject Pool Effects. Are treatment effects robust to different types of subject pools? In this essay we assess the possible subject pool effects of three different convenience samples that are popular in experimental research: student samples typically employed in the lab; the MTurk online subject pool; and the U.K. online subject pool mainatined by Nuffield

CESS.

Each of these three convenience samples have characteristics that might result in treatment effects that are distinct from each other. Laboratory experiments are typically conducted using undergraduate students. Student subject pools have characteristics (Druckman and Kam, 2011; Belot, Duch and Miller, 2015) that certainly distinguish themselves from non-student populations. And we have evidence that for certain treatment effects students and non-students differ (Belot, Duch and Miller, 2015; Alatas et al., 2009).

Online experiments often recruit subjects from crowd-sourcing platforms. MTurk is one of the most frequently used online crowd-sourcing platform for micro-tasks, called Human Intelligent Tasks (HITs). The registered MTurk workforce (called workers) browse the list of HITs, and can accept and complete HITs. Experimental research employing MTurk subjects has been widely published in leading journals in political science (e.g. Healy and Lenz, 2014; Malhotra and Margalit, 2014; Grose, Malhotra and Parks Van Houweling, 2015), economics (e.g. Horton, Rand and Zeckhauser, 2011; Olea and Strzalecki, 2014; Kuziemko et al., 2013) and sociology (e.g. Tsvetkova and Macy, 2014). Crowd-sourced subject pools are a convenience sample. Nevertheless, they are considerably more diverse than a typical lab student pool (Berinsky, Huber and Lenz, 2012). For instance, Kuziemko et al. (2013) show that the attrition rate for a multiple wave survey is lower for MTurk samples than some other experiments; they also contend that the MTurk sample is no less representative than conventional telephone surveys of the U.S. population.[1] And this diversity could result in treatment effects that are quite distinct than those we observe with other subject pools.

A third type of subject pool is the CESS U.K. Online subject pool that is strictly employed for online experiments following operational and ethical practices similar to those employed by the Nuffield CESS Lab. Subjects in these panels are recruited periodically to participate in online experiments. They are more likely to resemble a lab subject pool in that their participation in experiments is much more infrequent. On the other hand their

[1]Their comparison is to an unweighted CBS survey.

diversity makes them more similar to the MTurk subject pool. These distinct features of the CESS UK Online panel could distinguish their treatment effects from those of the other subject pools.

.

We propose to explore subject pool effects with different analytic strategies. This project makes comparisons between three different subject pools: student subjects from the CESS Lab subject pool, subjects from the CESS UK Online subject pool; and US subjects recruited from MTurk. Subjects from these different pools take part in identical online experiments in a virtual online lab. First, we simply compare the socio-demographic composition and preferences of the different sample pools. Second, we compare treatment effects across subject pools – if the treatment effects are quite similar it would be reasonable to conclude that subject pool effects are minimal. On the other hand if we find significant differences in treatment effects, this would suggest that the composition of subject pools affects treatment effects.

Mode Effects. Our second goal is to examine whether exclusion restrictions are violated when treatments for interactive experiments are administered online. It is frequently claimed that moving from the lab to the online environment undermines the internal validity of an experiment (e.g. Clifford and Jerit, 2014; Weigold, Weigold and Russell, 2013). Among other things, it is difficult to determine who is actually treated and the conditions under which subjects are making their decisions.

On the other hand, there clearly is evidence that the decisions of online subjects resembles those of subjects in the lab. Berinsky, Huber and Lenz (2012) reports that many of the established results from survey experiments using framing as a treatment can be replicated employing the MTurk subject pool (see also Crump, McDonnell and Gureckis, 2013; Rand, Greene and Nowak, 2012). Also, Grose, Malhotra and Parks Van Houweling (2015) maintain that MTurk subjects are more suitable for some types of treatments, such as reading lengthy

texts, than online national representative samples because "MTurk respondents are 'workers' who are accustomed to completing lengthy tasks in exchange for payment (p.734)."

Studies without incentivised treatment also examine the mode effects. For instance, Clifford and Jerit (2014) test several psychological measures in parallel surveys in lab and online using student subjects and find that although there are few differences in attention measures and social desirability effects, there is significant difference in political knowledge measure, suggesting that subjects consult with outside sources in online setting.

Concerns about exclusion restrictions are even more acute in the case of online interactive experiments in which subjects are making decisions in real-time. There are experiments in which online subjects play real-time synchronic games (e.g. Mason and Suri, 2012), but most studies in behavioural economics using online subjects do not use realtime interaction. Many studies use either one-shot game without feedback (e.g. Horton, Rand and Zeckhauser, 2011) or sequential interaction in which subjects leave a virtual lab and come back after several days (e.g. Tsvetkova and Macy, 2014).

Our focus will be on violations of exclusion restrictions for online interactive experiments. We measure the threats to internal validity associated with moving from a very controlled lab environment to an online experimental environment over which the experimenter has much less control. The ideal design, obviously, would be to observe the same subject making the same decisions in a controlled lab setting and in an online experimental setting. We try to approximate this ideal with our design.

Design. We implement very similar experimental protocols in the lab and online. These are public goods games in which subjects earn money performing real effort tasks; their earning are taxed and distributed to other group members (subjects are randomly assigned to groups of four or six depending on the experiment); and subjects have opportunities to lie about their earnings. First, subjects make the same strategic decisions. And to the extent possible we replicate the interactive feature of this decision making. Second, we observe the

subjects from the same subject pool (although not identical subjects) making these decisions in the lab and online. Third, we observe subjects from different types of online subject pools (that vary significantly in their diversity) making decisions in the same experiment.

If there are in fact significant mode effects then we would expect 1) to see overall differences in treatment effects across these different experimental modes; and 2) these differences in treatment effects should be particularly salient when we control for subject pool demographic characteristics. We should also be able to estimate subject pool effects since we observe, in particular, subjects from the CESS lab subject pool make decisions in the same experiment administered both in the traditional lab and online.

Subjects in the lab and online versions of the cheating experiments essentially play the same game although they make decisions in quite different environments. In the lab version of the experiment subjects are randomly assigned to groups; although anonymous, subjects know their group members are one of approximately 24 other participants in the lab facility; and the collective decisions of the subjects affect the payoffs of other group members in real time.

We implement three non-synchronic online versions of the public goods game employed by Beramendi and Duch (2014). Subjects are not interacting with each other in real time. Online subjects are assigned, through a substitution strategy, to groups that have already played the game. Payoffs to the online subjects are then determined by their choices and the choices that had already been made in previous lab experiments. And there is of course an absence of physical proximity online that might also affect the choices that are made by subjects.

A second set of online comparisons build on the initial lab experiments implemented by Duch and Solaz (2016). We implement synchronic online versions of Duch and Solaz (2016). These synchronic online experiments closely resemble the lab experiments including real-time decision making by groups of four subjects.

This essay is primarily concerned with whether there are mode effects associated with

the implementation of the these cheating experiments in online settings. Since we have implemented essentially identical experiments both in the lab and online we are able to address this question. And in particular we are interested in mode affects that are associated with experiments involving strategic interaction amongst subjects. As we point out above there have been some comparisons between experiments conducted with online subject pools and other types of subject pools, but these studies have not focused on experiments that involve strategic interaction amongst individuals.

Estimation Strategy. Our goal is to leverage, to the extent possible, variation in treatment effects across subject pools and experimental modes in order to understand the costs and benefits of moving outside of the traditional experimental lab. Table 1 summarizes our estimation strategy.

Table 1: Summary of Mode and Subject Pool Effects

Mode	Lab Subject Pool	Online Subject Pool	Subject Pool Characteristics
Lab BD	Lab-BD	NA	NA
Online Non-synchronic	Online Lab-BD	CESS Online-BD	Estimated
Online Non-synchronic (M-Turk)	NA	MTurk-BD	NA
Lab DS	Lab-DS	NA	NA
Online Synchronic (M-Turk)	NA	MTurk-DS	NA
Online Synchronic	Online Lab-DS	NA	Estimated
Mode Effect	Estimated	Estimated	

Note Mode and Subject Pool Effects.

Table 1 indicates that our strategy for identifying subject pool and mode effects is to conduct identical experimental protocols with different subject pools and modes. Our focus here is on experiments in which subjects make choices in an interactive environment in which the payoffs resulting from these decisions are conditional on the choices of other subjects who are deciding at exactly the same time. Certainly from the perspective of internal validity this represents the most challenging design for an online experiment. And its this interactive feature that is the most difficult to replicate in an online environment. Our focus will be comparing decision making and treatment effects across different types of subject pools and across experimental modes.

Amount of Cheating. Most important is comparing the decision to cheat in interactive decision making contexts. Accordingly, we estimate these subject pool and mode effects with cheating experiments that have been conducted both in the lab and online by Duch and Solaz (2016) and Beramendi and Duch (2014). These experiments all share a number of key characteristics.

Both the lab and online experiments have very similar set-ups. In both modes subjects are random assigned to groups (of four or six); they are informed of "deduction" rates, audit rates, and penalties for cheating; they are informed about how the group earnings are distributed to each of the four members; and they are given instructions on how to earn money in a real effort test (RET). Subjects are required to first undertake a real effort task and then secondly report their earned income from these tasks.

Comparisons between lab and online will be made for two versions of the public goods games – Duch and Solaz (2016) (no redistribution) and Beramendi and Duch (2014) (redistribution in terms of both taxation and expenditures). The key outcome metric for the games is variation in cheating.

Preferences. We are also interested in whether preferences differ in any significant fashion across subject pools – in particular preferences that might be significant confounders. Ac-

cordingly, the experiments gathered information on the characteristics of the online and lab subject pools. Subjects make decisions designed to recover basic preferences and personality characteristics: 1) a dictator game to measure other-regarding preferences; 2) a lottery game to measure risk preferences; and 3) and an integrity test to measure honesty.

Socio-economic Characteristics. These subject pools will of course vary in terms of socio-economic characteristics. We document these differences. In addition, we explore whether they play any significant role in accounting for variations in treatment effects across sample pools.

3 Experiments

3.1 Laboratory Experiments

Beramendi and Duch (2014)(BD). Two variations on public goods cheating games were conducted at the CESS experimental lab with University of Oxford students from the CESS student subject pool. One variation is based on Beramendi and Duch (2014). The experimental sessions were conducted from November 22 to December 3, 2013. The experiment consisted of five modules, two cheating modules (one with and one without auditing), dictator game, lottery game, and non-incentivised questionnaire. In the first three modules, we offered earnings in Experimental Currency Unit (ECU). The conversion rate is 300 ECUs to 1 British Pound. The cheating modules consist of ten rounds each. Prior to the cheating public goods game, participants are randomly assigned to groups of six and the composition of each group remains unchanged. Each round of these two cheating modules has two stages. In the first stage subjects perform RET to compute a series of two-number additions in one minute. Their Preliminary Gains depend on the number of correct answers, getting 150 ECUs for each correct answer. In the second stage, they are asked to report their gains. Depending on their rank of Preliminary Gains within the group,

subjects are assigned to one of the tax terciles. If their report is audited, they have to pay additional penalty – 50 percent of unreported gains are also deducted. The audit rate is 0 percent for the first tax module and is 10 percent for the second. Deductions applying to the six group members including the penalty are pooled and distributed amongst those members in accordance with a redistribution rate that varies by treatment. At the end of each round participants are informed of their Preliminary and Declared gains; whether these two amounts have been audited; the total tax contribution in the group; the amount they receive from the deductions in their group; and the earnings in the round. At the end of each tax module one round is chosen at random, and their earnings are based on their profit for that round. Participants are only informed of their earnings for each tax module at the end of the experiment.

Duch and Solaz (2016)(DS). The Duch and Solaz (2016) cheating public goods game resembles Beramendi and Duch (2014). After earning money in the same RET, subjects in the Duch and Solaz (2016) experiment play a modified version of a public goods game. They are randomly assigned to groups of four; there is a known deduction rate imposed on earned income; they report their earned income (with a known probability of being audited); the taxed income is then evenly distributed amongst members of the group. Both experiments also measured preferences and socio-demographic characteristics. The demographic variables were gender and income. Preference modules include a measure of trust, integrity, other-regarding preferences and risk aversion. The details of these games are provided in Duch and Solaz (2016) and Beramendi and Duch (2014) and on their Online Appendices.

3.2 Online Experiments

Our estimation strategy involves replicating these two interactive lab experiments online. There are two different versions of the online experiment – one corresponding to Beramendi and Duch (2014) and another replicating Duch and Solaz (2016). We also conduct the online

version with different subject pools as part of our effort to tease out the subject and mode effects.

Nonsynchronic online BS Experiment. The Beramendi and Duch (2014) research question called for implementing an identical version of the lab experiment with a much larger and more diverse online subject pool. In the lab experiment, subjects were assigned to groups and their final payments were conditional on the behaviour of the other subjects in the group. Replicating this group assignment and real-time feedback in an online setting is challenging particularly for a relatively large number of subjects. We solve this challenge by realtime matching of online subjects with the behaviour of lab subjects.

The sequence of the modules in the experiment is mostly the same as the lab experiment except for the number of rounds in each tax module. To prevent the subjects from losing attention to the experiment and lower the number of dropouts from the experiments, we reduce the number of round in each deduction module from ten to five. From the ten rounds of the cheating game, five with audits and five without audits, one round is randomly selected for the payment. In the data analysis of laboratory experiments, we only use the data of ten rounds, first five rounds of from each of two modules. In the cheating game, online participants played exactly the same real effort tasks, which are two-digits number additions for sixty seconds. Each correct answer is converted to 150 ECUs.

The wording used in the questions is closely assimilated to the one in the lab experiment. However, we trimmed down the instructions given to the subjects. In the laboratory experiments, the experimenter read the instruction aloud, and experimental subjects could not skip the instruction to proceed to the next screen until the experimenter finishes the recitation. Since this is not the case for online experiments, we tried to make sure that the experimental subjects receive the instructions by making the instructions shorter and clearer as well as by setting a minimum time to spend in each page of instruction before proceeding and adding more examples. The experiment was conducted on a virtual laboratory hosted

on Qualtrics, an online survey platform.

A critical component of the experimental design is assigning subjects to a deduction rate depending on how they perform in the real effort task compared to other subjects in their group. In this online experiment we recreate the group environment by randomly assigning subjects to one of groups in the lab, and replacing one of lab subjects in the group instantly in each round. After a subject performed RET, we matched each subject to a lab subject – the matching is based on the similarity of their Preliminary Gains. The online subject's deduction rate and after-deduction revenues are determined by her or his contribution and remaining members of the lab group. This matching is accomplished by linking a database on our web server to Qualtrics.[2] We set up an online database that stores the laboratory data on the experimental parameters (such as the audit rate and deduction scheme) as well as the real effort task performance and cheating rates in the laboratory sessions. Qualtrics survey software accesses the database through a web-service function. At the start of experiment, each participant was randomly assigned to one of the 11 laboratory groups, and then the information about the deduction scheme for the group is retrieved from the server and shown on the subject's instruction screen.

The matching of online subjects' behavior with previous lab behavior works in the following manner. In each round of the tax compliance game, the Qualtrics survey program sends the record of an online subject's Preliminary Gain of an online subject to the web server, then a program placed on the server retrieves the information of the group performance for the round and determines a lab subject to be replaced by fining a subject whose Preliminary Gain is the most approximate to the online subject's Preliminary Gain for the round. The program retrieves the data on deductions of other five subjects, and then passes the data on the total contributions, tax rate of online subjects, and audit information back to Qualtrics. The information is shown on the experimental screen before the online subject is prompted to declare the gain. With the Declared Gains of an online subject and information received

[2]A survey experiment using the similar technology is found in Boas and Hidalgo (2012).

from the web server, Qualtrics survey program produces the results of the round and online participants receive a feedback regarding the round's results, which is equivalent to the information the lab subjects have received.

We have run three online experiments using different subject pools. The first online experiment used 500 MTurk workers (MTurk-BD).[3] We published the HIT on October 3, 2014 and finished the data collection within a day. The second UK online experiment uses 150 subjects from the CESS subject pool (CESS Online-BD). The field period was February 20 to 27, 2015. The third experiment used the student subjects of University of Oxford (Online Lab-BD). The field period was March 18 to 20, 2016. For both general online subjects and student subjects we randomly invite three hundred registered subjects and the experimental website was open until the number of completion reaches the number we targeted. We used currency conversion rates different from the lab: 1000 ECUs is equal to one currency units, the British Pound Sterling for lab and UK online, and the US Dollar for MTurk. The experimental settings for UK is exactly the same as the US experiment other than the currency of the reward. And of course these results will be compared to the traditional lab results (Lab Student-BD).

Synchronic online DS Experiment In this particular set of experiments, we implement one of the Duch and Solaz (2016) lab experiments – specifically, the very simple baseline treatment: "wage rates" for the real effort task are identical for all subjects (in our online MTurk version subjects receive 15 cents USD (MTurk-DS); and the pooled deduction revenues are distributed evenly amongst all subjects in a group. The deduction treatment consists of ten rounds at a particular deduction rate. Prior to playing the cheating game, participants are randomly assigned to groups of four. The composition of each group remains unchanged for the deduction treatment module. Each round of the module is divided into two stages. In the first stage subjects perform a real effort task. This task consist of com-

[3]The worker qualifications we used were country of residence (United States), number of completed HITs (more than 500), and rate of previous HITs approved by requesters (99 percent or higher).

puting a series of additions of two 2-digit numbers in one minute. Their Preliminary Gains, which will be an endowment for the following stage, depend on how many correct answers they provide, getting 150 ECUs for each correct answer.[4] Once subjects have received information concerning their Preliminary Gains, participants are asked to declare these gains. A certain percentage (that depends on the treatment) of these Declared Gains is then deducted from their Preliminary Gains. These deductions are then evenly divided amongst the members of the group. Note that in each session the deduction rate is consistent. The deduction treatments implemented in this online experiment are the following: 10% and 30%.

At the end of each round participants are informed of their Preliminary and Declared gains; the amount they receive from the deductions in their group; and their earnings in the round. Subjects are paid at the end of the online experiment, and do not receive feedback about earnings until the end of the experiment. Participants receive visual instructions on their screens at the beginning of each module (unlike the lab experiments instructions are not read and explained aloud). After ten rounds of the cheating game, one round is chosen at random, and their earnings are based on their profit for that round. At the end of the experiment their earnings in ECUs are converted to USD at the exchange rate 1000ECUs = 1$ (this compares to 300ECUs = 1£in the Duch and Solaz (2016) lab experiment). Participants are then asked to answer a questionnaire that consists of an Integrity Test and a series of socio-demographic questions.

The synchronic online version of the Duch and Solaz (2016) was conducted with three different subject pools – again in an effort to tease out subject pool and mode effects: In the Online Lab-DS version, subjects from the Nuffield CESS Lab subject pool participated in the online cheating experiment. We also conducted an MTurk-DS version with U.S. subjects from the MTurk subject pool. All of these compared to the CESS lab subjects who participated in the Duch and Solaz (2016) lab experiments (Lab-DS).

[4]ECU: Experimental Currency Unit, which is converted to real money at a fixed rate.

4 Results: Lab versus Online Experiments Compared

4.1 Comparing Subject Pool Characteristics

Demographics We first compare the socio-demographic of the different subject pools. Concerns regarding subject pool effects build on the notion that these samples differ quite significantly in terms of their socio-demographic profiles. There is some evidence to this effect. Figure 1 indicates that the gender distribution of subjects in the lab and online are quite similar except for the UK online samples. Figure 2 confirms that there are age differences in the three subject pools. We know that MTurk workers tend to be younger than population survey samples (Berinsky, Huber and Lenz, 2012), and as we would expect, the undergraduate student subjects both in the lab and online are even younger on average. UK online subjects are similar to MTurk subjects. The age distributions for MTurk and UK online are significantly different from UK lab and online in both t-test and Wilcoxon rank sum test, but MTurk and UK online are not distinguishable.

Figure 1: Gender of Subjects

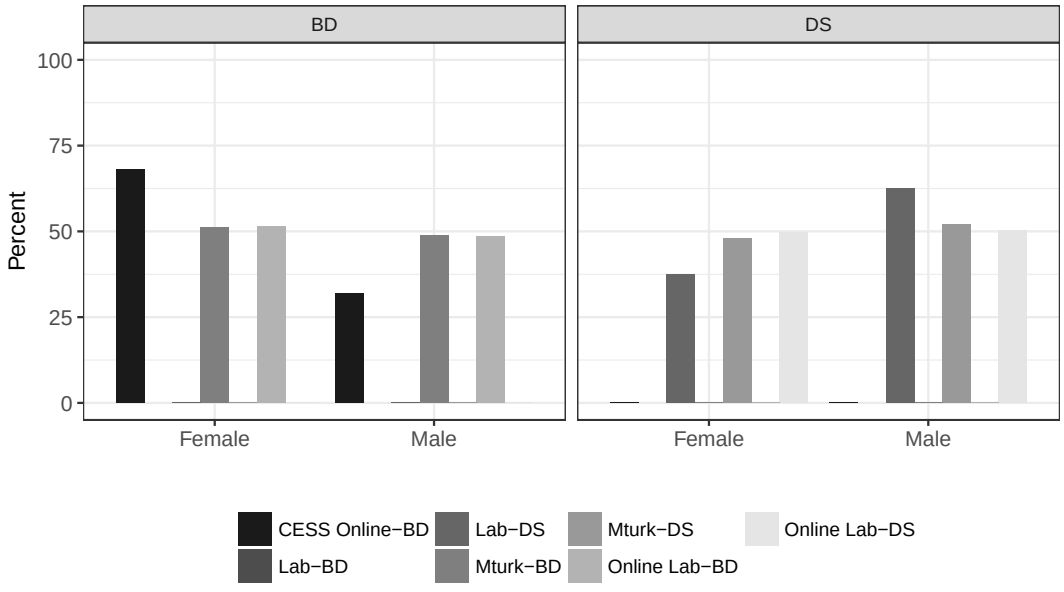
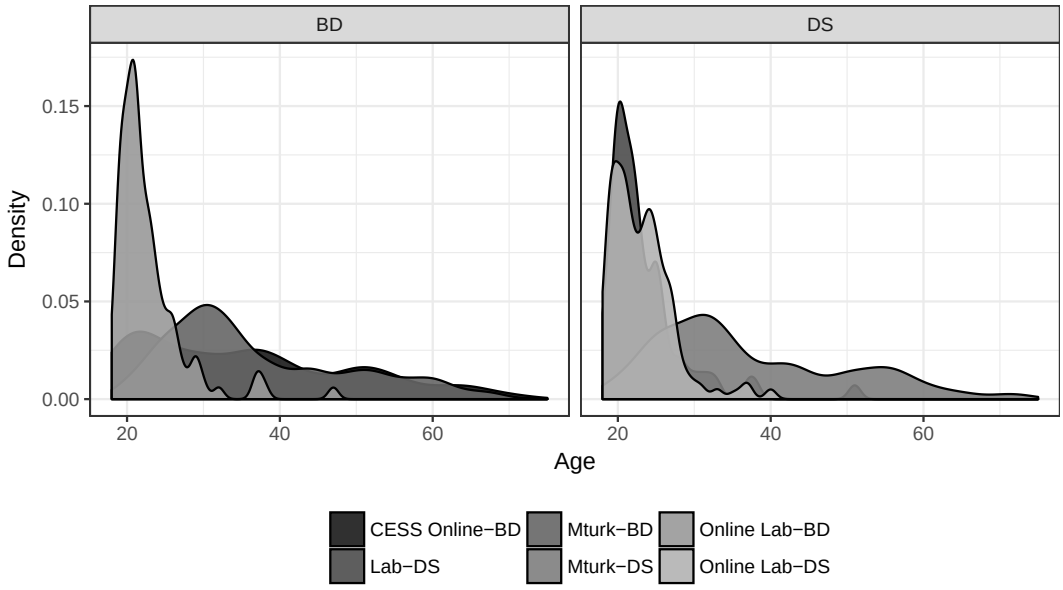


Figure 2: Age of Subjects

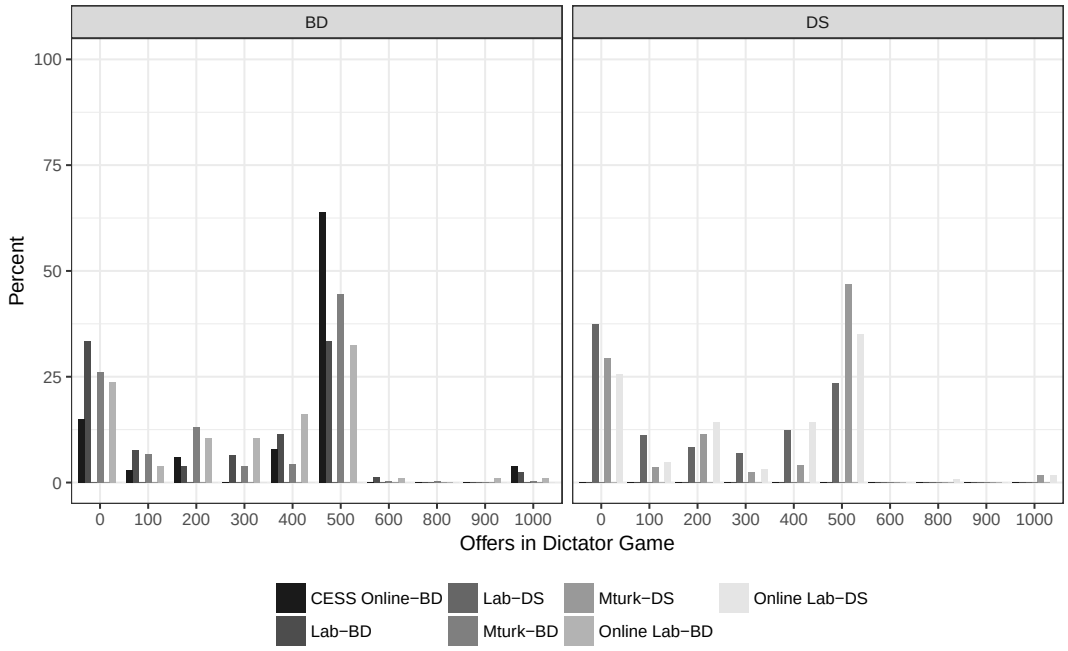


Preferences and personality. A more important concern though is that the subject pools differ with respect to fundamental preferences – this is the concern raised from the results reported in Belot, Duch and Miller (2015). As we pointed out earlier, both the online and lab experiments included the same set of behavioral measures for recovering underlying preferences that might be the source for heterogeneity in treatment effects. Comparisons of these different metrics across the different subject pools suggest no significant differences.

Other-regarding preferences are very similar across the different subject pools. We employ the classic Dictator Game described in the Online Appendix to measure other-regarding preferences. In both the lab and online versions of the Dictator Game subjects have an opportunity to split an endowment of 1000 ECUs between themselves and a recipient. Figure 4 describes the allocation of ECUs to the recipients. A large proportion of subjects either allocate nothing or a half of the endowment to the recipients. The average amount allocated to the recipient is 299 by student lab subjects; 302 by student online subjects; 283 by MTurk workers; and 408 by the UK Online subjects. Student and MTurk subjects are similar: in both t-test and Wilcox rank sum test, the difference between two groups is insignificant. In

contrast, the UK Online subjects are significantly more generous than the other two subject pools. This is confirmed by both t-test and Wilcox rank sum tests.[5]

Figure 3: Dictator Game

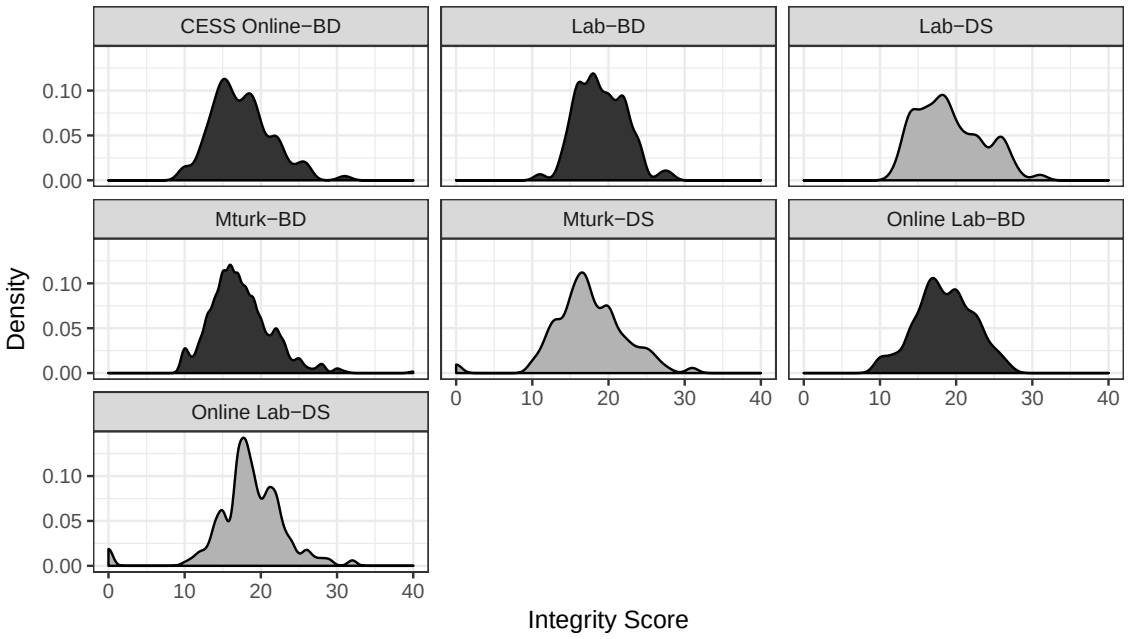


Integrity represents a personality trait that again may be the basis for heterogeneous treatment effects; particularly in our tax compliance experiments. The test consists of 10-item questionnaire described in Online Appendix. A summation of these ten scores results in an overall measure of integrity. Figure 4 shows the distribution of this total score which ranges from 10 (highest integrity) to 40 (lowest integrity). The distribution of scores suggests lower integrity (higher scores) for the student subjects. And in fact the average integrity score is significantly higher for lab subjects than online subjects. These differences in integrity between a student subject pool and more diverse subject pools are consistent with the student/non-student differences reported by Belot, Duch and Miller (2015).

A final source of treatment heterogeneity that might vary across the subject pools is risk preferences. The risk preference elicitation game described earlier is designed to measure

[5]The averages of scores in each measure from different pools and test results are presented in Online Appendix.

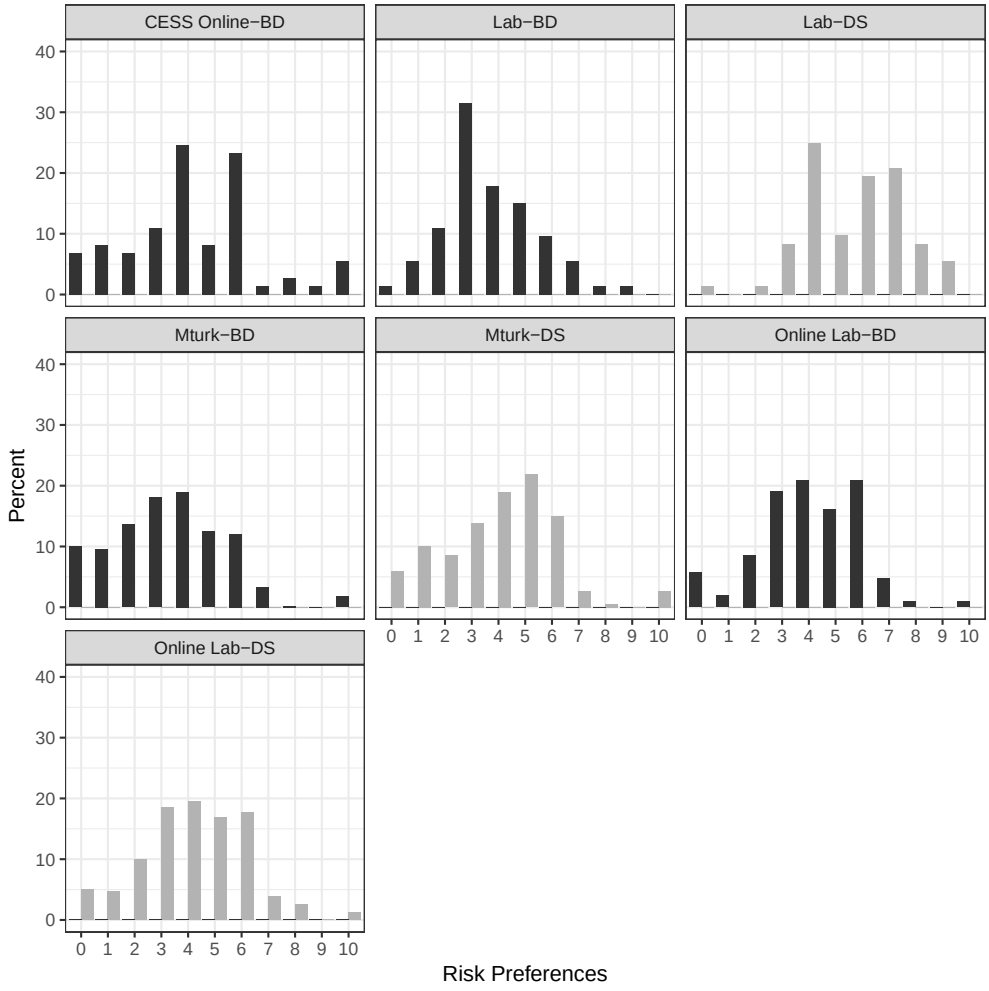
Figure 4: Integrity Test: Total Score



these preferences in both lab and CESS online subjects. The measure assumes transitive preference and monotonically non-decreasing utility in terms of monetary earnings. If a subject chooses Option B in a particular lottery, then in subsequent lotteries she should choose Option B. Violation of transitivity is often observed in these games. In this experiment, most lab and MTurk subjects reveals consistent preference: 7.6 percent of lab subjects and 9.8 MTurk subjects exhibit such inconstancy in the test. The rate is slightly higher for CESS online subjects, but the difference is not statically significant. On balance violation of transitivity represents less than 10 percent of subjects – we exclude these subjects from the analysis. However, UK CESS Online subjects exhibits surprisingly high inconsistency: 27 percent of subjects are inconsistent.

Figure 5 shows the distribution of risk preference after deleting these inconsistent results. The x-axis indicates the risk preference measured as the lottery choice switch. For example the score 4 means that a subject has switched their choice from Option A to B at Lottery 4. The score 0 represents the most risk seeking score while 10 represents the most risk aversive

Figure 5: Risk Preference



score.[6] Online subjects are slightly more risk seeking but on balance the three subject pools are quite similar with respect to risk preferences.

On balance, these subject pools do not differ dramatically on the preferences and personality measures we administered in these experiments. Potentially more problematic is the fact that, as we pointed out earlier, the lab and the online modes differ quite significantly. And these very different environments could affect treatment effects. We now explore this possibility. There are two key behavioural outcomes associated with the cheating experi-

[6]The score 10 is a logically inconsistent choice because a subject with this score has never switched their choice even if in Lottery 10 the Option B provide a higher earning with a probability of 1. There are eight such subjects among MTurk participants and four among UK Online.

ments: real effort performance and cheating.

Real Effort Performance. Recall that the tax compliance experiment required subjects to perform a real effort task that consisted of adding two randomly generated two-digit numbers in thirty seconds. Our expectation is that both lab and online subjects would treat the real effort tasks with similar levels of seriousness and they would register similar outcomes. Both subject pools are incentivised although the online subjects earn less and for many this is one of a large number of HITs that they perform. And as we pointed out earlier there is some question about how attentive online subjects are to these such tasks. Significant differences would raise question about the employment of real effort tasks with online experimental subjects.

Figure 6: Real Effort Task Performance

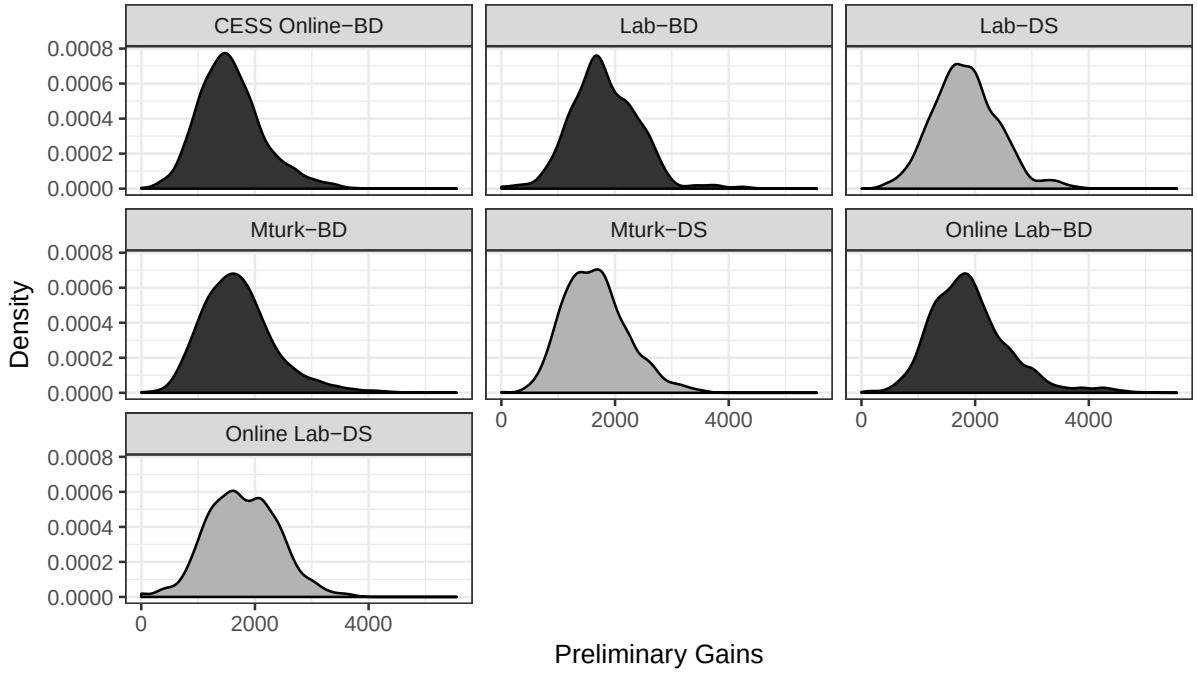


Figure 6 shows the distribution of outcomes for both lab and online subjects. Online subjects performed slightly lower than did the lab subjects: Average Preliminary Gains for lab student subjects is 1836 (an average of 12.2 correct answers), for online student subject is

1821 (12.1 correct answers), for MTurk subjects it is 1579 (10.5 correct answers); and for UK Online subjects it is 1425 (9.5 correct answers). Lab subjects who are university students are younger and, on average, better educated which might explain the higher average. MTurk subjects performance is higher than UK Online as they are "professionals" in such online work.

4.2 Cheating

Our outcome variable in these experiments is cheating. One might imagine, for example, that subjects in a lab experiment will be more hesitant about cheating given that they actually physically interact with other participants in the lab. There is little questioning the anonymity of online subjects and this might exaggerate cheating levels.

Beramendi and Duch (2014) Cheating. The design innovations that were necessary in order to replicate the cheating experiments online might also affect cheating. The online replication of Beramendi and Duch (2014) required the assignment of online subjects to groups of lab subjects who had already played the game – in fact, each online subject was substituted for a real lab subject who had already played the game.[7] Our expectation is that this change in the cheating experiment should not result in the online subjects complying with deduction rates differently than lab subjects. One of the goals of the comparisons we implement is to test this conjecture.

In the lab version of Beramendi and Duch (2014), the subjects are assigned to a six-member group at the start of the cheating experiment. The lab experiments took place months before the online experiment. In each of the ten rounds of the online experiment, online subjects are matched to one of the lab subjects based on the similarity of their performance on the real effort task. At the income reporting stage the online subjects

[7]There is no deception in how we implement the assignment of online subjects – they are told that they are being paired with subjects who had already made their decisions about how much of their income to declare.

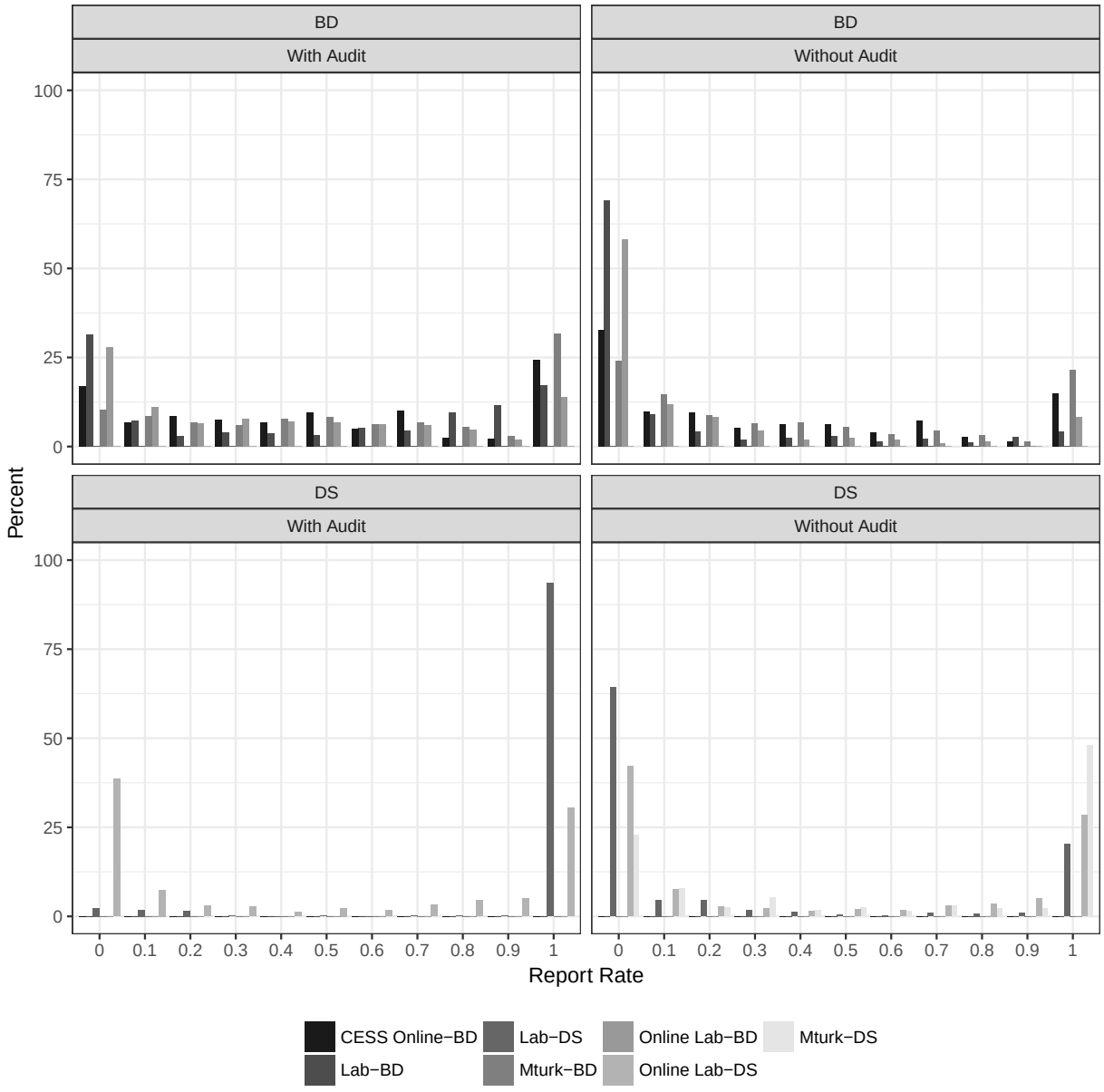
receives the identical information that was provided to their matched lab subject: they are informed of the deduction regime; they learn their deduction rate; and they are informed of the benefits regime that determines how the pooled deduction revenues are distributed to each member of the group.

If the online subjects happened to have the identical preference to their matched lab subjects, then the online subjects would make the exact same decision (this controls for treatment effects because the treatment effect is precisely the same in each pairing). In order to test this hypothesis, we matched the replacing subjects for each round with replaced subjects to make a comparable datasets of lab and online results. The matching of lab to online is 1-to-N; in other words, for each data point of lab data, there are "N" multiple points in the online data.

The top graphs in Figure 7 compare the Beramendi and Duch (2014) lab and online income report rates using the matched dataset. Reporting no income always maximized subjects' expected earnings in this experiment – in both audit rate treatments. The income report rate takes the value of one in a substantial number of online observations. With the positive audit rate case, about 16 percent of UK Online subjects and 12 percent of MTurk subjects reported the full income while 9 percent of lab student subjects reported full income. The proportion is even smaller for online student subject with 7 percent report rate. These proportions go down when there is no auditing of subjects' income; nevertheless, still 8 percent of UK Online subjects and 11 percent of MTurk subjects reported their full income while 2 percent of lab student subjects did so and 4 percent of online student subject did.

There appears to be a distinct subject pool effect here. The online subjects have similar cheating behavior – whether its the CESS UK Online subject pool or the U.S. MTurk subject pool. On the other hand subjects drawn from the CESS Lab subject pool – whether they play the cheating game online or in the lab – clearly are much more comfortable about cheating.

Figure 7: Comparison of Income Report Rate



Duch and Solaz (2016) Cheating. The synchronic online replications of Duch and Solaz (2016) closely resembled the lab versions in that online subjects were randomly assigned to groups of four and played the cheating game in real time with the other group members. The bottom graphs in Figure 7 compare the cheating behavior of CESS lab subjects in the lab, MTurk subjects online and CESS lab subjects online. One clear difference here is that

CESS lab subjects whether in the lab or online are clearly much less likely to fully report their income than is the case with MTurk subjects. In the zero audit condition, almost 50 percent of MTurk subjects report 100 percent of their income while only 10 percent of the CESS lab subjects, in either mode, behave similarly.

In the case of the Duch and Solaz (2016) experiment, there is a clear reticence on the part of MTurk subjects to cheat compared to the CESS lab subjects (either in the lab or online). This is certainly consistent with the Beramendi and Duch (2014) suggestion of a subject pool effect. A comparison of MTurk and CESS UK Online subjects would provide a more definitive assessment of this subject pool effect.

4.3 Comparing Treatment Effect

The "treatment" effects of interest in these experiments are developed and explored in detail elsewhere (Duch and Solaz, 2016; Duch, Laroze and Zakharov, 2017; Beramendi and Duch, 2014). Subjects in both experiments were assigned to similar deduction and audit treatments. Our general expectation is that those who perform better on the RET will cheat more; cheating will drop as deduction rates rise; and cheating will be lower when there is no auditing of income.

Table 2 reports results from the first set of models in which we regress the percent of income reported on dummy variables of deduction rate terciles, audit rates, and number of correct additions in the RET (performance). We include two dummy variables for deduction rates, Middle and Low, as well as a "No Audit" dummy variable. The baseline is the high deduction rate tercile and a 10 percent audit rate. For both online and lab models, all estimated coefficients are highly significant in the expected direction.

The are some notable difference between lab and online. First, the effect of no audit is larger for the lab subjects. For CESS lab subjects, playing the Beramendi and Duch (2014) cheating game in the lab, moving from the 10 percent audit to the no audit treatment results in an average decrease in reported income of 31 percent. The zero audit effect is an even

greater negative 71 percent for CESS lab subjects playing the Duch and Solaz (2016) cheating game. The online subjects are less responsive to the possibility to cheat without retribution.

Of particular interest is the Performance cheating effect proposed by (Duch and Solaz, 2016; Duch, Laroze and Zakharov, 2017): They argue that those who have higher than average performance on the RET will have a higher cheating proclivity. This is clearly the case across all of the subject pools and experimental modes. With the exception of Equation 1, the # of Additions variable is negative and significant.

Table 2: Effects of Tax Rates and Auditing

	CESS Online-BD	Mturk BD	Lab BD	Online Lab-BD	Mturk DS	Lab DS	Online Lab-DS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
# of Additions	−0.007	−0.005***	−0.026***	−0.019***	−0.013***	−0.017***	−0.011***
	(0.005)	(0.002)	(0.005)	(0.004)	(0.003)	(0.002)	(0.002)
Middle	0.085**	0.040**	−0.028	0.016		0.070***	
Tax Bracket	(0.039)	(0.017)	(0.036)	(0.030)		(0.020)	
Low	0.072	0.072***	−0.014	0.066*	−0.042**	0.126***	0.154***
Tax Bracket	(0.050)	(0.021)	(0.047)	(0.038)	(0.021)	(0.020)	(0.019)
No Audit	−0.137***	−0.168***	−0.310***	−0.194***		−0.706***	−0.026
	(0.023)	(0.011)	(0.024)	(0.020)		(0.017)	(0.019)
Constant	0.511***	0.598***	0.786***	0.589***	0.759***	1.089***	0.492***
	(0.085)	(0.035)	(0.081)	(0.061)	(0.033)	(0.033)	(0.032)
Observations	994	4,968	778	982	1,846	1,440	2,287
R^2	0.055	0.068	0.243	0.185	0.014	0.560	0.041
Adjusted R^2	0.052	0.067	0.239	0.182	0.013	0.559	0.040

*p<0.1; **p<0.05; ***p<0.01
Outcome variable: Percent of Income Reported

In Table 3 we compare mode effects for a more fully-specified model including the behavioral measures discussed in the previous sections. These results confirm the findings in Table 2 – the two treatments (performance and audit rate) are significant and in the same direction as in Table 2. The additional control variables in Table 3 have effect sizes that are roughly similar across modes and all are in the direction that we would expect: more

generous giving in the Dictator Game is negatively correlated with cheating; higher integrity (a low score) results in less cheating; and subjects who are more risk averse (a low score) are less likely to cheat.

Table 3: Effects of Behavioral Preference and Tax Compliance

	CESS Online-BD	Mturk BD	Lab BD	Online Lab-BD	Mturk DS	Lab DS	Online Lab-DS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
# of Additions	−0.001	−0.006***	−0.023***	−0.016***	−0.013***	−0.015***	−0.011***
	(0.006)	(0.002)	(0.005)	(0.004)	(0.003)	(0.002)	(0.002)
Middle	0.046	0.033*	−0.069*	0.009		0.028	
Tax Bracket	(0.043)	(0.017)	(0.036)	(0.030)		(0.021)	
Low	0.083	0.062***	−0.054	0.073*	−0.062***	0.088***	0.130***
Tax Bracket	(0.057)	(0.021)	(0.047)	(0.037)	(0.022)	(0.021)	(0.018)
No Audit	−0.184***	−0.178***	−0.318***	−0.205***		−0.705***	−0.038**
	(0.027)	(0.011)	(0.024)	(0.020)		(0.016)	(0.018)
Dictator Game	0.318***	0.311***	0.271***	0.158***	0.429***	0.323***	0.452***
Giving	(0.066)	(0.026)	(0.050)	(0.051)	(0.044)	(0.043)	(0.040)
Integrity Score	−0.116	−0.366***	−0.220*	−0.001	−0.010	0.115*	−0.176***
	(0.113)	(0.044)	(0.117)	(0.086)	(0.066)	(0.059)	(0.059)
Risk Preference	−0.054	−0.048*	−0.200***	−0.196***	0.107**	0.087*	−0.228***
	(0.055)	(0.026)	(0.070)	(0.062)	(0.053)	(0.045)	(0.047)
Constant	0.402***	0.637***	0.916***	0.609***	0.601***	0.944***	0.508***
	(0.113)	(0.039)	(0.115)	(0.084)	(0.045)	(0.051)	(0.042)
Observations	725	4,460	728	968	1,632	1,440	2,249
R^2	0.116	0.135	0.287	0.218	0.075	0.577	0.113
Adjusted R^2	0.107	0.134	0.281	0.212	0.072	0.575	0.111

*p<0.1; **p<0.05; ***p<0.01
Outcome variable: Percent of Income Reported

Respondents in all three of these experimental modes are randomly assigned to effectively the same treatments. One of our primary concerns in this essay is understanding whether treatment effects differ significantly across experiments. The audit treatment effect, for example, is clearly much stronger in the lab than online. We can explore whether this difference is the result of observed covariate differences in lab versus online subjects. Accordingly we balanced each of the samples by matching on the covariates of treatment assignments, such as deduction rate terciles, as well as the behavioral measures, such as other regarding preference and risk preference. We employ the method of covariate balancing propensity score method

(CBPS) proposed by Imai and Ratkovic (2014). We first calculate the propensity score in which the outcome is three subject sources, then weight samples based on the propensity score. By re-weighting, we can achieve the covariate balances across three subject pools. The model estimated in Table 3 was re-estimated employing these balanced samples. The results reported in Table 4, closely resemble those in Table 3. Note in particular that the audit treatment effect remains much higher for the CESS lab subjects. This suggests that there are likely unobservables that distinguish the Oxford undergraduate subjects from the CESS online and MTurk subject pools. As Belot, Duch and Miller (2015) suggest, for games in which other-regarding preferences likely matter, student subjects appear to be more *homo economicus* than non-student subject pools. And, of course, this difference might be exacerbated in the case of Oxford undergraduate subject pools which were employed in both this study and also in the study conducted by Belot, Duch and Miller (2015).

Detecting Mode Effects through Students Comparison In a further attempt to differentiate mode effect from subject pool effect, we combine the experimental data and then restrict our data to student subjects and estimate the models after matching on key covariates (King and Zeng, 2005). We adopt our matching strategy from King and Zeng (2005) and Ho et al. (2011). We first create a matched dataset using nearest neighbor matching. The treatment variable is the mode of experiment: lab versus online (the Lab dummy assumes a value of 1 for the lab experiments). And the covariates are the Age and Gender of subjects. The observations outside the convex hull of covariates in the laboratory subject pool are eliminated from the analysis. We also include a dummny variable for whether the study was (Duch and Solaz, 2016; Duch, Laroze and Zakharov, 2017) (in which case DS study assumes a value of 1) or Beramendi and Duch (2014) (DS Study assumes a value of 0).

Our interest here is to explore whether there are significant mode effects even after balancing demographic covariates. Since the matching results in subjects that are very

Table 4: Effects of Behavioral Preference and Tax Compliance (Covariate Balanced)

	CESS Online-BD	Mturk BD	Lab BD	Online Lab-BD	Mturk DS	Lab DS	Online Lab-DS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
# of Additions	−0.001	−0.006***	−0.023***	−0.016***	−0.013***	−0.015***	−0.011***
	(0.006)	(0.002)	(0.005)	(0.004)	(0.003)	(0.002)	(0.002)
Middle	0.046	0.033*	−0.069*	0.009		0.028	
Tax Bracket	(0.043)	(0.017)	(0.036)	(0.030)		(0.021)	
Low	0.083	0.062***	−0.054	0.073*	−0.062***	0.088***	0.130***
Tax Bracket	(0.057)	(0.021)	(0.047)	(0.037)	(0.022)	(0.021)	(0.018)
No Audit	−0.184***	−0.178***	−0.318***	−0.205***		−0.705***	−0.038**
	(0.027)	(0.011)	(0.024)	(0.020)		(0.016)	(0.018)
Dictator Game	0.318***	0.311***	0.271***	0.158***	0.429***	0.323***	0.452***
Giving	(0.066)	(0.026)	(0.050)	(0.051)	(0.044)	(0.043)	(0.040)
Integrity Score	−0.116	−0.366***	−0.220*	−0.001	−0.010	0.115*	−0.176***
	(0.113)	(0.044)	(0.117)	(0.086)	(0.066)	(0.059)	(0.059)
Risk Preference	−0.054	−0.048*	−0.200***	−0.196***	0.107**	0.087*	−0.228***
	(0.055)	(0.026)	(0.070)	(0.062)	(0.053)	(0.045)	(0.047)
Constant	0.402***	0.637***	0.916***	0.609***	0.601***	0.944***	0.508***
	(0.113)	(0.039)	(0.115)	(0.084)	(0.045)	(0.051)	(0.042)
Observations	725	4,460	728	968	1,632	1,440	2,249
R^2	0.116	0.135	0.287	0.218	0.075	0.577	0.113
Adjusted R^2	0.107	0.134	0.281	0.212	0.072	0.575	0.111

*p<0.1; **p<0.05; ***p<0.01
Outcome variable: Percent of Income Reported

similar on important covariates, and differences in the outcome variable are more likely to be the result of mode effects. Using the matched dataset, we regress the cheating measure used in the previous analyses, i.e., the average level of reporting in the No Audit treatment (Model 1) and the Audit treatment (Model 2) on the same set of explanatory variables plus the Lab and Study effect dummy variables. Table 5 reports the results for these two models. In both models, the effect of the Lab dummy variable is significant although in opposite directions. The lab mode seems to result in a more rational response to the audit treatment. When there is a non-zero audit probability subjects in the lab cheat less than subjects online. When there is no audit, subjects in the lab cheat more than subjects online. This seems to suggest that its not simply the composition of the lab subject pool that results in more homo-economicus behavior but that the mode itself seems to encourage rational and greedy

decisions.

Table 5: Effect of Modes

	Audit	No Audit
	(1)	(2)
# of Additions	-0.010^{***}	-0.017^{***}
	(0.001)	(0.001)
Lab	-0.365^{***}	0.199^{***}
	(0.015)	(0.014)
DS Study	0.018	0.150^{***}
	(0.011)	(0.010)
Dictator Game Giving	0.285^{***}	0.365^{***}
	(0.023)	(0.021)
Integrity Score	-0.466^{***}	-0.182^{***}
	(0.034)	(0.031)
Risk Preference	-0.094^{***}	-0.065^{***}
	(0.025)	(0.023)
Constant	1.070^{***}	0.337^{***}
	(0.029)	(0.027)
Observations	5,521	6,734
R^2	0.144	0.159
Adjusted R^2	0.143	0.159

Note: *p<0.1; **p<0.05; ***p<0.01. The dependent variable is the average reporting level of each subjects in Audit (Model 1) and No Audit (Model 2) treatments.

5 Discussion

Lab experiments can be an ideal vehicle for understanding cheating behavior. But moving from the experimental lab to online experiments with a much more diverse subject pool raises a number of challenges. In particular, in this particular case, it required scheduling real time interactions amongst online subjects which is in practice very difficult. In this essay we explore two different solutions. One non-synchronic solution was to exploit the existing experimental lab results from Beramendi and Duch (2014). We matched online subjects to lab subjects based on the similarity of their performance on real effort tasks. This allowed us to calculate deductions and redistribute deduction revenues to the online subjects. A second synchronic solution was to programme online experiments that allowed subjects to play the cheating game in groups of four in real time.

The goal of this article is to explore whether there are mode or subject pool effects associated with a lab versus online experiment of this nature. First we simply compare the characteristics of the three different subject pools. The student subject pool for the lab experiment was, as we expected, younger. There were no significant differences with respect to gender between lab and MTurk, but UK Online pool contained more female participants. Second, we compare the three subject pools in terms of personality traits and preferences. Here we do find some differences that are consistent with the findings of Belot, Duch and Miller (2015) who conduct similar comparisons of student and non-student subject pools. They find that students tend to be less trusting and exhibit lower levels of other-regarding preferences. We find that students subjects, compared to online subjects, show similar levels of trusting and other-regarding preferences, but score lower on a measure of integrity. On balance, though, the subject pools are reasonably similar in terms of personality traits and preferences.

Another issue we explored was whether the two subject pools perform differently on the real effort tasks that were a critical component of the cheating experiments. The younger and better educated lab subjects performed somewhat better on the real effort tasks. But

the differences were not dramatic suggesting that the online subject pool were attentive to the tasks hence were properly incentivised to perform well.

Finally, we were particularly interested in whether treatment effects varied significantly across mode or subject pool. Our multivariate analyses suggest that our three central treatment effects – performance, audit probabilities and deduction rate – were very similar across the different modes and subject pools. Consistent with the earlier findings Belot, Duch and Miller (2015), there is evidence that Lab subject pools (which in our case are entirely students) tend to be more utility maximizers than the online subject pool. For any given deduction or audit rate they cheat more than the online subjects. And as deduction rates rise or audit rates decline their cheating increases a faster rate than that of the online subject pool. Nevertheless, the directional effect of the deduction and audit treatments were similar across subject pools.

References

Alatas, Vivi, Lisa Cameron, Ananish Chaudhuri, Nisvan Erkal and Lata Gangadharan. 2009. "Subject Pool Effects in a Corruption Experiment: A Comparison of Indonesian Public Servants." *Experimental Economics* 12:113–132.

Arechar, Antonio A., Simon Gachter and Lucas Molleman. 2017. "Conducting Interactive Experiments Online." *Experimental Economics* .

Belot, Michele, Raymond Duch and Luis Miller. 2015. "A Comprehensive Comparison of Students and Non-students in Classic Experimental Games." *Journal of Economic Behavior and Organization* 113:26–33.

Beramendi, Pablo and Raymond M. Duch. 2014. "The Distributive Basis of Tax Compliance Introduction: Why Tax Compliance Matters.".

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–368.
URL: *http://pan.oxfordjournals.org/cgi/doi/10.1093/pan/mpr057*

Boas, Taylor and F. Daniel Hidalgo. 2012. "Fielding Complex Online Surveys using rApache and Qualtrics." *The Political Methodologist* 20(2):21–26.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen and Hang Wu. 2016. "Evaluating replicability of laboratory experiments in economics." *Science* .

Clifford, Scott and Jennifer Jerit. 2014. "Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies." *Journal of Experimental Political Science* 1(2):120–131.
URL: *http://scottaclifford.com/wp-content/uploads/2012/02/Mode_CliffordJerit_Final1.pdf*

Coppock, Alexander. 2016. "Generalizing from Survey Experiments Conducted on Mechanical Turk : A Replication Approach.".

Crump, Matthew J C, John V McDonnell and Todd M Gureckis. 2013. "Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research." *PloS one* 8(3):e57410.
URL: *http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3596391&tool=pmcentrez&rendertype*

Druckman, James N. and Cindy D. Kam. 2011. Students as Experimental Participants. In *Cambridge Handbook of Experimental Political Science*, ed. James N. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupia. Cambridge: Cambridge University Press pp. 41–57.
URL: *http://ebooks.cambridge.org/ref/id/CBO9780511921452A017*

Duch, Raymond, Denise Laroze and Alexei Zakharov. 2017. "Is Cheating a National Pastime? Experimental Evidence." Nuffield Centre for Experimental Social Sciences Working Paper.

Duch, Raymond and Hector Solaz. 2016. "Who Cheats: Experimental Evidence from the Lab." Working Paper. Centre for Experimental Social Sciences, Nuffield College, University of Oxford.

Grose, Christian R., Neil Malhotra and Robert Parks Van Houweling. 2015. "Explaining Explanations: How Legislators Explain their Policy Positions and How Citizens React." *American Journal of Political Science* 59(3):724–743.
URL: *http://doi.wiley.com/10.1111/ajps.12164*

Healy, Andrew and Gabriel S. Lenz. 2014. "Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy." *American Journal of Political Science* 58(1):31–47.
URL: *http://doi.wiley.com/10.1111/ajps.12053*

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. 2011. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Retrieved from: http://gking.harvard.edu/matchit/docs/matchit.pdf." *Journal of Statistical Software* 42(8).
URL: *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.329.3100*

Horton, John J., David G. Rand and Richard J. Zeckhauser. 2011. "The online laboratory: conducting experiments in a real labor market." *Experimental Economics* 14(3):399–425.
URL: *http://link.springer.com/10.1007/s10683-011-9273-9*

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society* 76(1):243–263.

King, Gary and Langche Zeng. 2005. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2):131–159.
URL: *http://pan.oxfordjournals.org/cgi/doi/10.1093/pan/mpj004*

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez and Stefanie Stantcheva. 2013. "How Elastic are Preferences for Redistribution? Evidence from Randomized Survey Experiments." National Bureau of Economic Research Working Paper 1886.

Malhotra, Neil and Yotam Margalit. 2014. "Expectation Setting and Retrospective Voting." *The Journal of Politics* 76(04):1000–1016.
URL: *http://www.journals.cambridge.org/abstract_S0022381614000577*

Maniadis, Zacharias, Fabio Tufano and John A. List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *American Economic Review* 104(1):277–90.
URL: *http://www.aeaweb.org/articles?id=10.1257/aer.104.1.277*

Mason, Winter and Siddharth Suri. 2012. "Conducting behavioral research on Amazon's Mechanical Turk." *Behavior research methods* 44(1):1–23.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/21717266*

McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis* 10(4):325–342.
URL: *http://pan.oxfordjournals.org/content/10/4/325\nhttp://pan.oxfordjournals.org/content/10/4/3*

Morton, Rebecca and Kenneth Williams. 2009. *From Nature to the Lab: Experimental Political Science and the Study of Causality*. Cambridge University Press.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman and Jeremy Freese. 2016. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(02):109–138.
URL: *http://journals.cambridge.org/abstract_S2052263015000196*

Olea, Jose Luis Montiel and Tomasz Strzalecki. 2014. "Axiomatization and Measurement of Quasi-Hyperbolic Dicounting." *The Quarterly Journal of Economics* pp. 1449–1499.
URL: *http://qje.oxfordjournals.org/content/early/2014/05/12/qje.qju017.abstract*

Rand, David G, Joshua D Greene and Martin a Nowak. 2012. "Spontaneous giving and calculated greed." *Nature* 489(7416):427–30.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/22996558*

Tsvetkova, Milena and Michael W Macy. 2014. "The social contagion of generosity." *PloS one* 9(2):e87275.
URL: *http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3923723&tool=pmcentrez&rendertype*

Weigold, Arne, Ingrid K Weigold and Elizabeth J Russell. 2013. "Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods." *Psychological methods* 18(1):53–70.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/23477606*